

# Identifying, naming and interoperating data in a phenotyping platform network: the **GOOD**, the **BAD** and the **UGLY**.

**Romain David**<sup>1</sup>, Jean-Eudes Hollebecq<sup>1</sup>, Llorenç Cabrera-Bosquet<sup>2</sup>, Hanna Ćwiek-Kupczyńska<sup>3</sup>, François Tardieu<sup>2</sup> and Pascal Neveu<sup>1</sup>; Contact: [romain.david@inra.fr](mailto:romain.david@inra.fr)

<sup>1</sup> MISTEA, INRA, Montpellier SupAgro, Université de Montpellier, Montpellier, France <sup>2</sup> LEPSE, INRA, Montpellier SupAgro, Université de Montpellier, Montpellier, France, <sup>3</sup> Institute of Plant Genetics, Polish Academy of Sciences, Poznań, Poland

The EPPN<sup>2020</sup> is a research project funded by Horizon 2020 Programme of the EU that will provide European public and private scientific sectors with access to a wide range of state-of-the-art plant phenotyping installations, techniques and methods. Specifically, **EPPN<sup>2020</sup> includes access to 31 plant phenotyping installations**, and joint research activities to *develop novel technologies and methods for environmental and plant measurements*.

Here we present the **results of the discussions** of the 2019 annual project meeting to adopt community-approved architectural choices. It focuses on **persistent identification of data** and real objects, the **naming of variables** and the **priorities for increasing interoperability** among phenotyping installations. We describe the main elements to prioritize (the good) in order to enhance Findable, Accessible, Interoperable and Reusable (FAIR) quality for each data management system with a *pragmatic concern for all partners*.

## Focus on identification...

The plant phenotyping community gathers different actors with various means and practices. Among all the recommendations, the community requests identification methods (including the use of ontologies) *compatible with the 'local' pre-existing ones*. The identification scheme being adopted is based on Uniform Resource Identifiers (URIs) with independant left and right parts for each identifier. (based on ePID recommendations)

[https://www.pidconsortium.eu/?page\\_id=122](https://www.pidconsortium.eu/?page_id=122)



### The **GOOD**

- Use **non ambiguous and persistent identifier**
- Use **minimal information**, get rid of everything that may change.
- Require **external identifier (B2HANDLE, e-PIC...)** if your authority is not persistent enough.
- Provide **multiple output format** (.txt, .html, .csv, etc.) and link them together, so the user will have the choice.
- Integrate/upgrade already existing identifiers in a URI.
- Use persistent-URL with 303 redirect status.
- Associate creation date to help understanding.



### The **BAD**

- Unnecessary metadata in the identifier
- Ownership and other information that are likely to change over time, prefer nature of the resource
- Unnecessary long identifiers with too much semantic
- Entirely opaque identifier
- Files extension in the URI (no .extension in the URI)
- Query (no "?" in the URI)
- Misleading characters such as O and 0 or I and l, etc.
- URI that are not the best way to identify the object you are looking at



### The **GOOD**

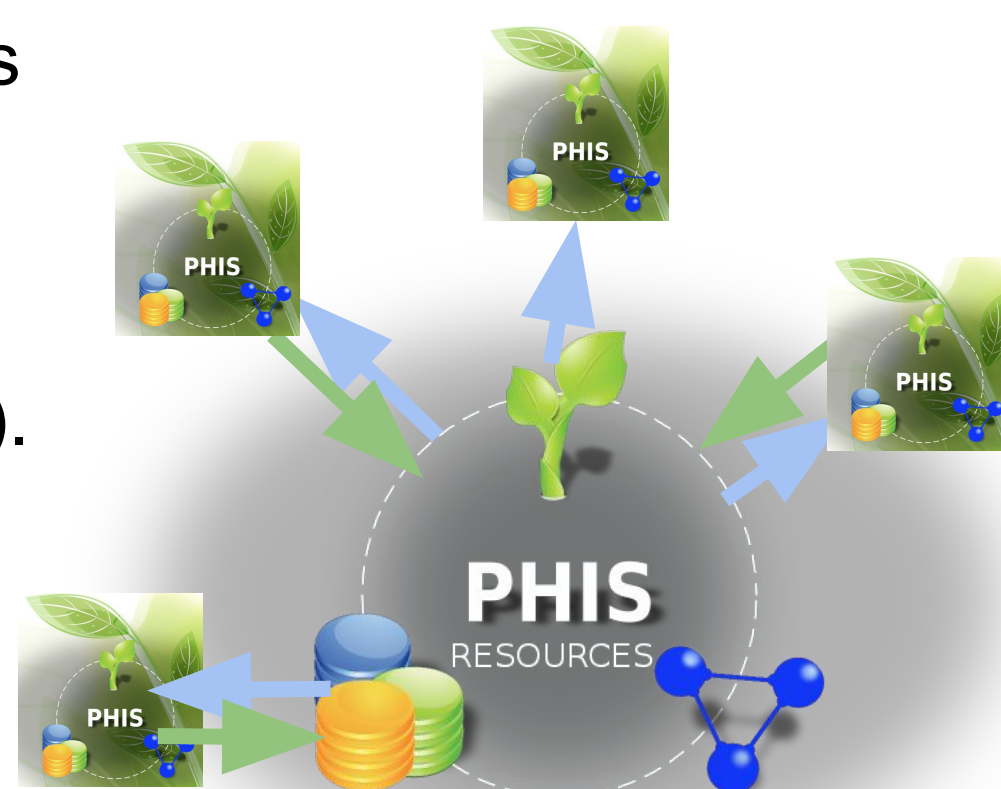
- Look for **"reference ontologies"**, first in the dataweb stacks
- Be careful of needs and implementation capacities to manage ontology links on the long term
- Look for concepts related to **your phenotyping experiments available in "application ontologies"** in your disciplinary domain first, before creating new ones.
- Ontologies should **never be developed isolated**. Use SKOS to link as much data as possible to reference ontologies first, and to trade/application ontologies with "exact match" then "close match" SKOS predicates
- If you need a **new concept, try to do it in concertation with the larger community** (as far as possible)



### The **BAD**

- To create an ontology before prospecting an existing one
- To create an ontology without a community approving process
- To give a URI for an ontology with date or version in the persistent link
- To use first a species specific ontology before considering concepts from general and recommended plant ontologies
- To use approximated data type
- To refer to approximated data concept in your specialized ontology

A common architecture for identifiers and variable names is being built in order to enable a first level of interoperation between Phenotyping Hybrid Informations Systems (PHIS). All instances are connected to a PHIS Resource center **using ontologies** and enabling sharing between each instance of PHIS.



### The **UGLY**

**Next challenges** that need to be addressed by the EPPN<sup>2020</sup> community are related with:

- the **partial reuse** of pre-existing ontologies,
- the **persistence of long-term access** to data,
- interoperation **between all potential users** of the phenotyping data.